**§1. Основные понятия математической статистики**

**1. Генеральная совокупность –** множество объектов, объединенных по какому-либо признаку или признакам (мужчины какой-либо области, достигшие призывного возраста, популяции птиц данного вида, обитающие в конкретном регионе, зерна пшеницы, собранные с данного поля и так далее).

**2**. **Выборка –** часть генеральной совокупности, отобранная для исследования.

**3. Суть выборочного метода в математической статистике:** по выборке судят о генеральной совокупности. Выборочные числовые характеристики являются приближенными значениями (оценками) генеральных числовых характеристик.

**4. Репрезентативность** (представительность). Выборка всегда с некоторой ошибкой характеризует генеральную совокупность. Например, выборочное среднее есть величина случайная, то есть она может принимать то или иное значение, а генеральное среднее – величина не случайная, является постоянной. Поэтому значение выборочного среднего рассеяны с некоторой дисперсией вокруг генерального среднего. Чем больше выборка (то есть, чем больше объектов отобрано из генеральной совокупности), тем меньше ошибка репрезентивности, то есть, тем лучше характеризует выборка генеральную совокупность.

**5. Вариационный ряд.** После того как объекты отобраны из генеральной совокупности, над ними производят измерения и получают тем самым экспериментальные данные. Эти данные нужно упорядочить, чтобы их легче было подвергнуть статистической обработке и, кроме того, некоторые характеристики признака выявляются уже после упорядочения данных. Упорядочения осуществляется путем построения вариационного ряда, который представляет собой таблицу с двумя графами. В первой графе указываются значения признака, а во второй сколько раз эти значения встретились в выборке (то есть частоты этих значений). Обычно значения признака перечисляются в возрастающем порядке. Если признак представляет собой непрерывную величину, то строят интервальный вариационный ряд, то есть указывают не отдельные значения признака, а разбивают весь интервал варьирования признака на классы – серию интервалов, границы которых указывают в соответствующей графе. А во второй графе указывают, сколько значений попало в тот или иной класс.

**Пример построения вариационного ряда.** Пусть дана выборка (размер зерен пшеницы в мм ):

5.41 5.54 5.49 5.35 5.40 5.26 5.50 5.46

5.41 5.55 5.31 5.45 5.54 5.32 5.52 5.39

5.62 5.40 5.23 5.45 5.47 5.40 5.42 5.45

5.58 5.50 5.36 5.44 5.50 5.37 5.47 5.50

Имеется, как видно, 32 варианта. То есть, выборочных значений. Длина зерен – величина непрерывная. Поэтому нужно строить интервальный вариационный ряд.

Для этого определим пределы варьирования признака в выборке, то есть найдем минимальную и максимальную варианты. Они подчеркнуты: Хmin=5.23, Хmax=5.62. Теперь определим величину классового интервала по эмпирической формуле:

$$i=\frac{X\_{max}-X\_{min}}{51gn}$$

i – величина классового интервала

n – объем выборки (число вариант в выборке).

$$i=\frac{5.61-5.23}{51g32}=\frac{0.39}{7.52}≅0.05$$

Первый интервал выберем таким, чтобы наименьшая варианта попала примерно в середину класса. Например, выберем границы первого класса такими: 5.20–5.25. Далее, прибавляя к правой границе очередного класса по 0.05 получим все другие классы.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Хi | 5.205.25 | 5.255.30 | 5.305.35 | 5.355.40 | 5.405.45 | 5.455.50 | 5.505.55 | 5.555.60 | 5.605.65 |
| fi | 1 | 1 | 2.5 | 5 | 7 | 7 | 5 | 1.5 | 1 |

Теперь каждую варианту нужно соотнести к соответствующему классу и подсчитать, сколько вариант попало в каждый класс. Если варианта попадает на границу двух классов, то в каждой из них заносят по 0,5. Заполнение второй графы таблицы завершает построение вариационного ряда

**6. Гистограмма.** Для наглядности сгруппированные данные можно представить в виде гистограммы. Гистограмма – это фигура, образованная столбиками, высота которых равна числу вариант, попавших в соответствующий класс (то есть частоте), а основанием столбика является соответствующий классовый интервал. Ниже приведена гистограмма, построенная по данным предыдущего примера.

fi

 Хi

Из гистограммы сразу видно, где сосредоточена основная масса данных. Кроме того, огибающая столбики гистограммы является прообразом функции плотности вероятности случайной величины, выборочные значения которой использовались для построения гистограммы. Поэтому в данном примере можно предположить, что случайная величина Х-размера зерна пшеницы, имеет нормальное распределение.

**7. Понятие о статистической оценке**. На практике в большинстве случаев всю генеральную совокупность подвергнуть исследованию невозможно (да это и не нужно).Ведь чтобы узнать вкус супа, нет надобности съедать всю кастрюлю. Достаточно взять пробу из генеральной совокупности – выборку, числовые характеристики которой являются приближенными значениями числовых характеристик генеральной совокупности. Приближенное значение генерального параметра, вычисленное на основании экспериментальных данных, называется **статистической оценкой**. Таким образом, статистика может дать только приближенное значение генерального параметра. Насколько приближенное? Чтобы статистическая оценка была достаточно близка к истинному значению числовой характеристики генеральной совокупности, она должна соответствовать нескольким требованиям.

8. **Требования к статистическим оценкам**. Статистическая оценка есть величина случайная, так как ее значения зависят от того, какие объекты из генеральной совокупности попали в выборку. Чтобы оценка правильно отражала генеральные параметры, к ней предъявляются требования **несмещенности, эффективности и состоятельности.**

**9.** **Несмещенность** оценки означает, математическое ожидание (среднее значение) оценки равно оцениваемому параметру при любом объеме выборки:

,

где  - статистическая оценка;  - истинное значение генерального параметра.

Это значит, что значения оценки должны иметь среднее значение, равное оцениваемому параметру.

**10.** Статистическая оценка должна быть **эффективной,** то есть иметь при заданном объеме выборки наименьшую дисперсию. Это значит, что значения оценки должны лежать вблизи истинного значения.

**11.** Статистическая оценка должна быть **состоятельной,** то есть стремиться по вероятности к оцениваемому параметру при . Это значит, что оценка должна быть такой, при увеличении объема выборки, n, точность приближения возрастала.

**12.** Несмещенной, эффективной и состоятельной **оценкой математического ожидания** нормально распределенной случайной величины является **среднее арифметическое,** определяемое формулой:

$$\overbar{X}=\frac{x\_{1}+x\_{2}}{n}=\frac{\sum\_{}^{}x\_{i}}{n}$$

Где X̅ - среднее арифметическое; *х*1, *х*2,…*х*i - значения признака (варианты) в выборке; n - объем выборки.

Многие характеристики биологических объектов – рост, вес, концентрация метаболитов, активности ферментов и другие имеют распределение близкое к нормальному. Поэтому на практике для оценки среднего значения случайных величин в биологических исследованиях чаще всего используется среднее арифметическое.

**13.** Несмещенной, эффективной и состоятельной **оценкой дисперсии** нормально распределенной случайной величины является $S^{2}$, величина которой вычисляется по формуле:

$$S^{2}=\frac{\sum\_{}^{}\left(x\_{i}-x̅\right)^{2}}{n-1}=\frac{\left(x\_{1}-x̅\right)^{2}+\left(x\_{2}-x̅\right)^{2}+…+\left(x\_{n}-x̅\right)^{2}}{n-1}$$

$$s=\sqrt{\frac{\sum\_{}^{}\left(x\_{i}-x̅\right)^{2}}{n-1}}$$

Величина S является оценкой стандартного отклонения случайной величины.

Оценки, приведенные выше, называются **точечными**. Они представляют собой точки на числовой оси. Однако точечная оценка не содержит в себе информации о том, насколько близка эта оценка к истинному значению оцениваемого параметра. Этот недостаток устраняется использованием интервальных оценок.

**14.** **Доверительный интервал**. Доверительным интервалом называются границы, между которыми с некоторой вероятностью (называемые доверительной) находится истинное значение оцениваемого параметра. Записывается доверительный интервал в виде (А, В). Из определения следует что

$p\left(A\leq θ\leq B\right)=$*β*

где β – доверительная вероятность; $θ$ - оцениваемый параметр, А и В – границы доверительного интервала.

В статистических таблицах часто вместо доверительной вероятности указывают **уровень значимости**. Связь между ними определяется равенством:

β+α=1,

где β – доверительная вероятность; α – уровень значимости. Например, если β = 0.95, то α = 1-0.95 = 0.05.

**15.** **Вычисление доверительного интервала для математического ожидания случайной величины, имеющей нормальное распределение.**

Если стандартное отключение известно, то для вычисления доверительного интервала используются следующие формулы:

$$(A,B)=(\overbar{x}-∆, \overbar{x}+∆)$$

$$∆=t∙\frac{σ}{\sqrt{n}}, $$

∆ - величина, равная половине доверительного интервала

Σ – стандартное отклонение

t – значение функции распределения вероятности стандартного

 нормального распределения, соответствующее доверительной

 вероятности,

n – объем выборки

Величину t находят по таблице функции распределения нормального распределения для соответствующей доверительной вероятности. В биологических исследованиях обычно используют три уровня доверительной вероятности: 0.95, 0.99 и 0.999.

**Пример**. Дано: $\overbar{x}=10.0, σ=2.0, n=25$

Построить доверительный интервал для $β=0.95$

Решение:

 $∆=t∙\frac{σ}{\sqrt{n}}=1.96∙\frac{2.0}{\sqrt{25}}=0.78$

$$A=\overbar{x}-∆=10.0-0.78=9.22$$

$$B=\overbar{x}+∆=10.0+0.78=10.78$$

$$\left(A,B\right)=(9.22,10.78)$$

Если дисперсия случайной величины не известна, то коэффициент t находят по таблице Стьюдента. Величина t в этом случае определяется доверительной вероятностью и числом степеней свободы при построении доверительного интервала для математического ожидания вычисляется по формуле:

$$K=n-1,$$

 где K – число степеней свободы; n – объем выборки.

**Пример.** Возьмем те же данные, что и в предыдущем примере, но значение $σ$ будем считать не точным, а приближенным, то есть будем считать, что оно вычислено по выборке. Тогда имеем:

$$∆=t∙\frac{S}{\sqrt{n}}=2.06∙\frac{20}{\sqrt{25}}=0.824≈0.82$$

Здесь величина 2.06 найдена по таблице Стьюдента для К = 24 и α = 0.95.

Таким образом, (A,B) = (9.18, 10.82).

Мы видим, что границы доверительного интервала раздвинулись. Это связано с тем, что точное значение дисперсии заменено приближенным, неопределенность возросла, расширился и доверительный интервал.

**Контрольные вопросы и задания**

1. Что такое генеральная совокупность?
2. Что такое выборка?
3. Суть выборочного метода.
4. Понятие о репрезентативности.
5. Понятие о статистической оценке.
6. Требования к статистическим оценкам.
7. Понятие о доверительном интервале.
8. Дана выборка (рост студентов в см): 166, 171, 167, 156, 167, 166, 167, 173, 169, 176, 168, 163, 169, 164, 171, 169, 161, 167, 167, 171, 168, 162, 174, 173, 173, 165, 167, 172, 176, 174, 170, 164, 166, 168, 172, 163, 170, 156, 154, 165, 177, 183, 163, 170.

Построить вариационный ряд и гистограмму, вычислить X, S и доверительный интервал для генеральной средней при β=0.95.

1. Дана выборка (длина зерен пшеницы в мм): 5.41, 5.55, 5.31, 5.45, 5,54, 5.32, 5.52, 5.39, 5.62, 5.40, 5.23, 5.45, 5.47, 5.40, 5.42, 5.53, 5.44, 5.41, 5.47, 5.51, 5.48, 5.40, 4.44, 5.60, 5.57, 5.47, 5.45, 5.38, 5.33, 5.64, 5.46, 5.56, 5.53, 5.53, 5.60, 5.47, 5.47, 5.47, 5.48, 5.58, 5.51, 5.52, 5.51, 5.47, 5.52, 5.48, 5.34, 5.59, 5.45, 5.34, 5.33, 5.41.

Построить вариационный ряд и гистограмму. Вычислить $\overbar{x},S и$доверительный интервал для генеральной средней, β=0,99.

1. Дана выборка (длительность сердечного цикла в сек.):

0.68, 0.52, 0.58, 0.59, 0.56, 0.74, 0.54, 0.72, 0.76, 0.76, 0.74, 0.79, 0.66, 0.84, 0.85, 0.81, 0.59, 0.71, 0.74, 0.65, 0.96, 0.83, 0.78, 0.94, 0.93, 0.62, 0.60, 0.57, 0.69, 0.56, 0.45, 0.55, 0.66, 0.49, 0.71, 0.68, 0.65, 0.66, 0.72, 0.92, 0.74, 0.66, 0.61, 0.61, 0.60, 0.67, 0.60, 0.69, 0.63, 0.80.

Построить вариационный ряд и гистограмму. Вычислить $\overbar{x},S и$доверительный интервал для генеральной средней при β=0,999.

1. Дана выборка (вес мозга крыс в мг):

1072, 1098, 986, 1127, 1122, 1143, 1273, 1070, 1190, 1165, 1074, 993, 1071, 1032, 952, 880, 978, 1025, 1111, 946, 1004, 945, 1108, 1169, 1126, 1137, 1268, 1004, 960, 928, 1021, 1073, 1118, 1126, 1085, 1001, 1040, 1070, 1029, 1049, 1000, 1008, 978, 1094, 877, 1168, 1098, 1100, 1031, 1058, 1086, 1004, 1088, 1035, 994, 997, 1028.

Построить вариационный ряд и гистограмму. Вычислить $\overbar{x},S и$доверительный интервал для генеральной средней при β=0,99.

1. Дана выборка (содержание митохондриального белка в мозге крыс в мг/г мозга): 17.1, 15.5, 14.6, 9.9, 11.9, 23.0, 14.6, 19.7, 13.8, 17.0, 12.3, 14.6, 13.2, 6.9, 5.8, 17.9, 15.0, 11.7, 14.4, 8.8, 14.3, 19.0, 9.8, 15.3, 183, 68, 12.3, 28.2, 13.8, 18.9, 17.7, 20.4, 22.6.

Построить вариационный ряд и гистограмму. Вычислить $\overbar{x},S и$доверительный интервал для генеральной средней при β=0,99.

1. Дана выборка (количество белка в синаптосомах мозга крыс в мг/г мозга): 9.9, 10.4, 6.8, 8.6, 5.5, 5.1, 3.9, 3.9, 7.0, 4.3, 5.2, 4.4, 6.4, 2.1, 1.3, 8.8, 3.4, 2.2, 7.4, 6.2, 5.0, 5.0, 10.8, 7.0, 5.0, 10.8, 7.0, 5.0, 3.2, 7.8, 5.2, 2.1, 4.4, 5.6, 4.8, 8.4, 3.1, 1.3, 2.6, 2.8, 6.5, 5.7, 4.3, 7.5, 2.1.

Построить вариационный ряд и гистограмму. Вычислить $\overbar{x},S и$доверительный интервал для генеральной средней при β=0,99.

14. Дана выборка (количество белка в водорастворимой фракции мозга крыс в мг/г мозга): 65.0, 50.0, 58.0, 50.0, 55.1, 60.0, 63.3, 42.2, 37.4, 63.4, 4.48, 7.42, 3.57, 57.0, 63.3, 53.6, 58.5, 55.1, 42.2, 37.4, 40.6, 48.8, 53.6, 42.0, 42.0, 57.0, 43.9, 55.1, 47.0, 52.0, 37.4, 48.7, 42.2, 70.0, 73.0, 35.7, 35.7, 43.8, 40.6, 30.9, 34.1, 37.4, 34.1, 42.2.

Построить вариационный ряд и гистограмму. Вычислить *X,S*  и доверительный интервал для генеральной средней при *β =* 0,95.

15. Дана выборка (длина зерен пшеницы в мм):

5.38, 5.47, 5.51, 5.30, 5.40, 5.40, 5.28, 5.48, 5.46, 5.51, 5.29, 5.42, 5.69, 5.60, 5.45, 5.38, 5.46. 5.52, 5.43, 5.18, 5.61, 5.36, 5.39, 5.44, 5.41, 5.41, 5.55, 5.31, 5.45, 5.54 5.32, 5.52, 5.39, 5.62, 5.40, 5.23, 5.45, 5.47, 5.40, 5.42, 5.45, 5.32; 5.44, 5.37.

Построить вариационный ряд и гистограмму. Вычислить *X,S* и доверительный интервал для генеральной средней при *β* = 0,95.

16. Дана выборка (содержание белка в грубой митохондриальной фракции в мг/г мозга):

61.0, 51.0, 70.0, 59.0, 63.3, 63.3, 51.0, 51.0, 49.0, 53.5, 56.0, 51.0, 41.0, 40.0, 40.0, 63.0, 39.0, 51.1, 46.3, 53.6, 65.8, 65.8, 48.7, 51.0, 51.0, 58.5, 53.6, 47.7, 57.0, 41.4, 41.4, 80.3, 90.0, 36.6, 41.5, 48.8, 65.6, 48.7, 63.5, 48.8, 58.3, 63.3.

Построить вариационный ряд и гистограмму. Вычислить *X,S* и доверительный интервал для генеральной средней при *β =* 0,95.